RECOGNIZING CALL-CENTER SPEECH USING MODELS TRAINED FROM OTHER DOMAINS

Erica Bernstein, Don McAllaster, Larry Gillick, Barbara Peskin

Dragon Systems Inc./Lernout & Hauspie
320 Nevada St.
Newton, MA – USA

ABSTRACT

In this paper, we introduce a new conversational speech task – recognizing call-center speech – using data collected from Dragon's own technical support line. We compare performance of models trained from conversational telephone speech (the Switchboard corpus) and models trained from predominantly read, microphone speech, and report on a series of experiments focusing on adapting the microphone speech models to the telephone channel and conversational task. We also discuss the importance of task-specific language model data. We benchmark our test set by comparing the performance of our 1998 Switchboard Evaluation system to that of our simpler call-center system.

1. INTRODUCTION

In this paper we investigate what happens when we take models trained for other tasks/domains and apply them to a new task for which we have no transcribed data: recognition of telephone calls to Dragon Systems' technical support line. The goal of the study was not to produce a highly optimized multi-pass system as we have for Switchboard evaluations [1], but rather to use existing technology to produce a fast, deployable system, providing transcripts in close to real time. This is in contrast to other tasks such as the IBM Voice-Mail transcription task where the recognition is done off-line by a multi-pass system [2, 3]. While call-center speech is a "real", goal-oriented task, unlike the more artificial Switchboard task, we felt we could leverage our experience with both Switchboard and the somewhat less contrived CallHome in order to recognize the call-center speech.

The target domain uses data collected from telephone conversations between members of Dragon's technical support staff and customers using one of our products. The content of the calls is fairly narrow, mainly focusing on questions concerning the interactions between various software and/or hardware. The technical support agent was recorded over a high-quality headset while the customers' speech was recorded from the telephone line. This paper focuses on experiments done recognizing the telephone speech of the customer. A customer test set was created with 20 callers – 14 male and 6 female – totalling approximately 10,000 words, representing an hour of speech.

Since the recognition task has no transcribed training data, we were forced to investigate the portability of models. Our experiments used models trained on data from other domains, where the primary focus was investigating how much we suffer when we use models mismatched in speaking style and from mismatched channels.

For the acoustic data, we compare performance obtained using models trained from the Switchboard corpus of conversational telephone speech and models trained on a corpus composed of an assortment of high-quality microphone-speech data, including the Wall Street Journal, selected Broadcast News, and in-house data. We note the microphone-speech is primarily read speech.

For language modeling, we combined data from three corpora: we used the Switchboard corpus to capture the conversational nature of the calls, Broadcast News for more general English, and e-mails to our technical support center for enriching the vocabulary with Dragon product names and computer jargon.

In the sections that follow, we provide a description of our models together with baseline recognition results in Section 2, and we detail a series of adaptation experiments in Section 3. We look into the role of the technical support e-mails as part of the language model in Section 4, and in Section 5 we explore the inherent difficulty of the task/test set by comparing the performance of our 1998 Switchboard evaluation system on this test with that of our simpler call-center system.

2. BASELINE MODELS AND RESULTS

2.1 Acoustic models

We have built parallel sets of acoustic models from equal amounts of data from the conversational Switchboard corpus and the microphone-speech corpus described above. The latter was downsampled to 8kHz in order to recognize telephone speech. We were particularly interested in the amount of degradation we would suffer using models trained from only downsampled microphone data, where there is a mismatch not only in the channel but in the speaking style as well.

We initially built two sets of acoustic models, one from 170 hours of Switchboard data (SWB) and one from 170 hours of the downsampled microphone-speech (HQMic). Both models were built using the same recipe: i.e. speaker-independent, (unwarped) triphone models with the same phoneme set, same feature set, same channel normalization and (after downsampling) same signal processing. The general mixture models had the same number of output distributions and the same bound on the number of Gaussians per mixture model.

2.2 Lexicon and language model

We used three language models for this task: one trained from nearly 3 million words of Switchboard training texts (SWB), another from approximately 145 million words of Broadcast News data (BN), and the third from ½ million words of the technical support e-mails (TS). The language model used in the experiments is a trigram language model, interpolated at the probability level from these three sources, according to the formula:

$$\text{Prob} = 0.53\ \text{SWB} + 0.24\ \text{BN} + 0.23\ \text{TS}\ .$$

We used a 50k-word vocabulary composed of ~28k words from Switchboard, 500 new words from the tech support e-mails, and the rest from Broadcast News. The test set has a 1% out-of-vocabulary (OOV) rate with the TS component and 2% OOV rate without the TS component. We discuss the OOV rate in Section 4, where we quantify the contribution of the technical support e-mails.

2.3 Baseline results

To establish a baseline, we run a simple "call-center" system that uses no adaptation, no warping, and only a single recognition pass. All recognition uses the language model (LM) described above and a fast decoding protocol, running between 1 and 2 times real-time on a PIII-450.

As seen in Table 1, for speaker-independent recognition, we began with a word error rate (WER) of 50.4% for the models trained from Switchboard data and 56.9% for the models trained from downsampled microphone-speech. The 6.5% difference is presumably due in large part to the channel mismatch and/or the mismatch in speaking style.

In our experience with the conversational telephone tasks, Switchboard and CallHome, the error rates are much lower using models trained from the same 170 hours of Switchboard data as the SWB models used here. We therefore wondered how much of the degradation in performance on call-center data could be explained by the compromises we made to our system and how much could be attributed to the task/test set. We discuss this more in Section 5.

Models	WER (%)
HQMic	56.9
SWB	50.4

Table 1: Baseline word error rates for Customer Technical Support Test Set.

3. ADAPTATION EXPERIMENTS

We were interested in techniques we could use to recover the loss of performance we suffered when using the models trained from read microphone data, particularly techniques involving adaptation to the telephone channel. Given the lack of transcribed training data we have for the target task, we were especially interested in the performance difference between supervised and unsupervised techniques. We were also interested in the quantity of data (transcribed or not) required to see improvements from our techniques.

We first exposed the microphone models (HQMic) to general telephone speech by performing supervised Baum-Welch (BW) adaptation to the 170 hours of Switchboard data. This brought the WER for the HQMic models down to 51.3%, within a point of the Switchboard-trained base model. This is not particularly surprising since both the SWB models and the adapted HQMic models have "seen" the full Switchboard acoustic training data together with the correct transcripts. On the other hand, when we adapt the HQMic models using unsupervised BW adaptation, based on Switchboard transcripts obtained by a fast errorful (~50% WER) recognition pass using SWB-trained models, we see a WER of 55.4%. This only amounts to a 1.5 point improvement absolute.

We were interested to see if the limited amount of task-specific technical support data we have could be used effectively to sensitize the microphone-speech models to the telephone channel in place of the Switchboard training data. Because of the limited amount of available data, we used regression-based (MLLR-style) adaptation [4, 5], rather than Baum-Welch. As adaptation data, we used the test set itself, jack-knifing through the data, leaving out the data for the speaker we were testing on so as to adapt to the task rather than the individual speaker. Using unsupervised adaptation, this resulted in a WER of 54.6%, which is better than adapting (unsupervised BW) to the 170 hours of Switchboard data and much cheaper. If we perform the jack-knifing experiment in supervised fashion, we achieve a word error rate of 52.7%. These results are summarized in Table 2. It is worth noting that we performed the same experiment starting from the Switchboard-trained models and observed no improvements, reinforcing the idea that we are adapting the microphone models to the channel and/or speaking style.

Adaptation data	Supervised	Unsupervised
170 hours Switchboard	51.3	55.4
1 hour technical support	52.7	54.6

Table 2: Word error rates for adapting HQMic Models to channel/task.

Since transcribing data is so expensive, we investigated how much we gain performing supervised regression-based adaptation as we increase the amount of transcribed data. These results are summarized in Table 3. For each experiment, we limited the amount of data per speaker so that the sum over all speakers came to the total minutes in the first column, and we used the same jack-knifing technique as for the experiments above. The models used in the last row are the same models used in the last row of Table 2. We first note that with as little as 10 minutes of transcribed data we obtain a 54.3% WER, noticeably better than adapting (unsupervised) to the 170 hours of Switchboard data. It is also worth noting that we attain our "steady-state" performance after 40 minutes of adaptation on transcribed data. For the unsupervised experiments, the total improvement of 56.9% to 54.6% WER is achieved after exposing the models to 20 minutes of speech.

	Supervised	Unsupervised
no adaptation	56.9	56.9
10 minutes	54.3	55.3
20 minutes	53.0	54.6
40 minutes	52.7	---
1 hour	52.7	54.6

Table 3: Word error rates for adapting HQMic models to varying amounts of task-specific technical support data.

We were curious to see whether gains from adapting to channel/task would persist after adapting to the speaker. We adapted to the speaker by adapting to the recognizer's output on the test data and then re-recognizing. We performed this speaker adaptation to 4 models: to the microphone models (HQMic), to the models obtained by doing the jack-knifing experiments on the 1 hour of technical support data in both supervised and unsupervised forms (JK-sup, JK-unsup), and to the Switchboard-trained models (SWB). The first column of Table 4 gives the baseline results for the models, as reported in Table 1 for HQMic and SWB and in Table 2 for JK-unsup and JK-sup. The remaining columns give word error rates if we only perform speaker normalization ("warping") [6, 7] but not adaptation, or if we both warp and adapt to the speaker.

Model	Baseline	Warp at test	Warp+adapt to speaker
HQMic	56.9	56.1	53.6
JK unsup	54.6	54.3	52.1
JK sup	52.7	52.5	50.8
SWB	50.4	49.2	47.9

Table 4: Adapting to speaker: word error rates with no warp / no adapt, warp only, and warp+adapt.

Comparing the baseline results for HQMic and JK-sup, we see a 4.2 point gain from adapting to the channel, assuming transcripts are available. After adapting to the speaker, we still see a 2.8 point improvement. The 50.8% WER for the JK-sup models is almost as good as the baseline results for the SWB-trained models. However after adapting the SWB-trained models to the speaker, they out-perform the speaker-adapted JK-sup models. It is worth remarking generally, that – although the size of the differences may narrow – within each column and row, strict ordering of the word error rates is preserved. Comparing the baseline results to the adapted results for HQMic and JK-unsup, we found that adapting to the channel without using the transcriptions was still beneficial even after adapting to the speaker. We note that neither of the channel-adapted models benefit from warping as much as we would expect.

4. LANGUAGE MODEL EXPERIMENTS

As described in Section 2.2, for this task we used three language models interpolated at the probability level. As seen in Table 5, without the technical support component of the language model, the performance is much worse, even though the OOV rate is still fairly low. The impact of the technical support data on the OOV rate may be deceptively small because of choices we made in text normalization and tokenization, i.e. in defining what constitutes a word. For example, the test set has many computer-specific multi-word phrases and numerous number/letter combinations which we retained as separate words rather than forming task-specific compounds. These phrases are often composed of fairly common words, such as "c colon backslash", "3 point O", "two sixty six", "windows ninety five", and "rich text format". A significant contribution of the TS component may therefore be in learning new ways to connect these common words, a hypothesis supported by the last row of Table 5, where we add in the technical support data, but only with unigram counts, without higher-order *n*-grams. This improves performance over the pure SWB+BN system, but we gain as much again by adding in the trigrams, most likely due to the structure of the multi-word computer jargon. We used the SWB acoustic models for these experiments.

Language model	WER	OOV
SWB + BN + TS (trigrams)	50.4	1 %
SWB + BN (no tech support LM)	56.9	2 %
SWB + BN + TS(1) (unigrams for TS)	53.9	1 %

Table 5. Word error rates and OOV rates using language models with TS component, without TS, and with only TS-unigrams.

5. BENCHMARKING THE TASK

As mentioned in Section 2.3 when discussing the baseline results, the 50.4% word error rate for the models trained from Switchboard data (SWB) was higher than we expected. In order to determine how much of the error could be attributed to the task/test and how much was due to the compromises we made to our system, we ran this test set through the 1998 version of our Switchboard system (eval'98) used in NIST's Hub 5 evaluations. All tests used the same SWB+BN+TS language model.

In Table 6 we see that if we use our eval'98 system on the customer test set, the error rate can be brought down to 37.7%. Though somewhat worse than typical Switchboard evaluation results, this is in line with results we have seen, for example, on the English CallHome task.

The eval'98 system uses a multi-pass protocol. It uses models trained from warped data and uses warping to test speakers (but no adaptation) in the first pass. Subsequent passes adapt to the speaker. In Table 7 we show the comparable results for the call-center SWB system. To make the fairest comparison, the results in Table 7 are taken from the last row of Table 4, where we warped at test time and then adapted to the speaker by adapting to test recognition and re-recognizing, as described in Section 3.

First pass (warp only)	41.9
Final Pass (adapt)	37.7

Table 6. Word error rates for 1998 Switchboard Eval System on Customer test set.

SWB (warp only)	49.2
SWB (adapt)	47.9

Table 7. Word error rates for SWB models used for call-center system.

In the initial pass, the eval'98 system has a 41.9% WER, compared to the SWB models' 49.2%, a gap of more than 7 points. By the final pass, the gap widens to 10.2 points.

The differences between the two experiments that were largely responsible for the performance difference in the initial pass are that the eval'98 system uses warped training data, has larger acoustic models (including more state models and more nodes per phoneme), and uses channel normalization by conversation side (rather than by utterance). The eval'98 system also uses a different phoneme set and different signal-processing parameters, although we have verified that the different phoneme set does not significantly affect performance. We expect that the differences in signal-processing features is a small benefit for the eval'98 system, but a relatively minor one compared to the differences already cited. In the final pass, the eval'98 system profits from a more elaborate multi-pass adaptation protocol. It also uses looser thresholds, running at roughly 20 times real-time compared to the nearly real-time call-center system.

6. FUTURE WORK

This task has provided us with valuable insights into the creation of deployable "real world" recognition systems. We plan on doing further experiments to pinpoint the performance differences between our optimized eval'98 system and the simpler call-center system. The goal is to narrow the gap between the two systems while working toward fast on-line transcriptions for the call-center system. We also plan more channel/task adaptation experiments where we compare short regression-based adaptation using Switchboard data to the jack-knifing experiment done with the technical support data. We are also interested in exploring the relationship between WER and the success of the calls and/or the emotionality of the speaker. Finally, while this paper focused on the telephone speech of the customer's side of the calls, we also have the opportunity to investigate the conversational speech of the technical support agents, who were recorded over high-quality microphones, and plan a series of experiments studying speaking style and channel mismatch conditions for the agent.

REFERENCES

1. Peskin, B., *et al*., "Improvements in recognition of conversational telephone speech," *Proceedings of ICASSP-99*, Phoenix, March 1999.
2. Padmanabhan, M., Saon, G., Basu, S., Huang, J., Zweig, G., "Recent improvements in voicemail transcription," *Proceedings of Eurospeech'99*, Budapest, Hungary, September 1999.
3. Huang, J., Padmanabhan, M., "A study of adaptation techniques on a voicemail transcription task," *Proceedings of Eurospeech'99*, Budapest, Hungary, September 1999.
4. Nagesha, V., Gillick, L., "Studies in transformation based adaptation," *Proc. ICASSP-97*, Munich, April 1997.
5. Peskin, B., *et al*., "Progress in recognizing conversational telephone speech," *Proc. ICASSP-97*, Munich, April 1997.
6. Wegmann, S., *et al*., "Speaker normalization on conversational telephone speech," *Proc. ICASSP-96*, Atlanta, May 1996.
7. Peskin, B., *et al*., "Improvements in Switchboard recognition and topic identification," *Proc. ICASSP-96*, Atlanta, May 1996.